Missing Value Treatment using Effective Optimization on Data from Multiple Social Media

[1] Sukhman Kaur, [2]Neeraj Sharma , [3]Kawaljeet Singh
[1]Research Scholar, [2]Professor, [3]Director
[1,2]Department of Computer Science, [3]University Computer Center
Punjabi University, Patiala, Punjab, India

Abstract
Missing value are broad in numerous genuine applications. Missing value imputation and in addition treatment is vital on the grounds that the skipping of missing value based records can harm the general results. For instance, if the client conclusions about information leak in India are fetched from social media then the client having hidden personal information can be covered in missing records. Such records cannot be skipped because of the privacy concerns of the users and therefore missing value imputation should be implemented on such records. In this research work, random forest approach for missing value imputation is devised and implemented on the different types of social media like youtube, twitter, tumblr.

Keyword- *Social media, Random Forest Approach, Missing value, Missing value imputation*

INTRODUCTION

The growth of social media over the last decade has revolutionized the way individuals interact and industries conduct business. Individuals produce data at an unprecedented rate by interacting, sharing, and consuming content through social media. Understanding and processing this new type of data to glean actionable patterns presents challenges and opportunities for interdisciplinary research and tool development. [4]

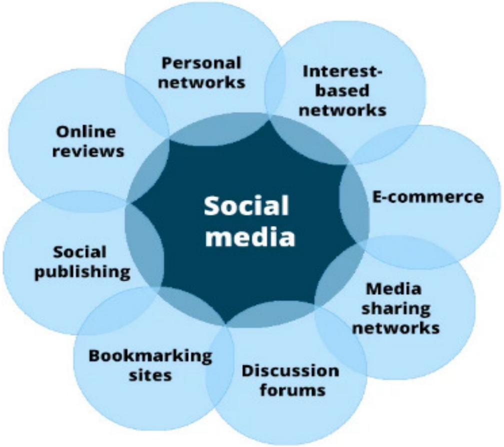


Figure 1. Taxonomy of Social Media

Social Media Mining integrates civil media, mutual incorporate analysis, and announcement mining to laid at one feet a enjoyable and accordant platform for students, practitioners, researchers, and duty managers to know the facts and potentials of social media mining. [9]

Missing data are a common issue in most clear research spaces, for instance, Social Media Analysis, Satellite Data, GPS Data, Biology, Medicine or Climatic Science. They can rise out of different sources, for instance, mistreating of tests, low banner to-change degree, estimation goof, non-response or eradicated interesting admiration. Missing values make it difficult for analysts to carry out statistics evaluation in data. Basically three kinds of issues are normally related to missing value

1) Loss of efficiency.
2) Complications in handling and analyzing the data.
3) Bias resulting from differences between missing and complete data.

Statistician categorized missing data into three categories as:
(a) Missing not at Random (MNAR).
(b) Missing at Random (MAR)
(c) Missing completely at Random (MCAR)

As indicated by Rubin, MAR is to be a condition in which the likelihood that information are missing depends just on the watched information yet on the missing information, in the wake of controlling for watched information. Missing totally indiscriminately (MCAR) is the possibility of a record having a missing estimation of the trait however it doesn't relies upon the missing information or the watched information. Not missing at Random (NMAR) is the likelihood of a record containing missing estimation of field that relies upon the estimation of clothing[28]. The following are common methods for missing value imputation:

* **Mean:** The mean of the observed values for that variable
* **Substitution:** The substitution based filling or imputation is based on the neighborhood based imputation without fragmentation of the dataset and this approach is not time consuming and not optimized in nature.
* **Hot deck:** The random selected of values is done in the hot deck whereby the randomly selected values based on the probability is done. Such probability based outcome is not effective in all types of datasets and more complexity aware.
* **Cold deck:** The cold deck approach is based on the similarity based approach in which the most similar values nearby to the occurrence levels are evaluated and finally imputed in the dataset.
* **Regression:** The predicted value obtained by regressing the missing variable on at variance variables.
* **Stochastic regression:** The integration or amalgamation of the values is done in the stochastic regression that is the hybrid approach of regression and randomizations.
* **Interpolation and extrapolation:** The case scenario of estimation and inference based imputation is done in the interpolation based outcome that is less commonly used because of its dependency in the dataset and not fit for all types of data.

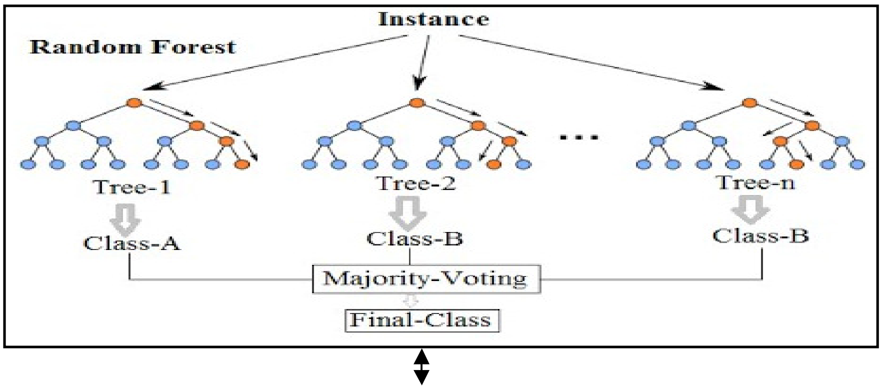
Problem definition and related work

Problem Definition and Research Gaps while identification of the problem is that in the traditional strategies and algorithms to deal with the missing qualities include removal of the records from real dataset. If such strategy is used, the results can be affected. In traditional research work of missing qualities imputation, the statistical evaluation using variance of existing values are used which can be improve using effectual algorithm.

Mean based imputation is having evident focus on the classical mean in which the traditional averages of all the values are done. Such process is done to have the mean value and that mean value is further fed to the imputation perspectives.
The generation of linear regression curve and prediction based on that curve is done on scenarios of regression and prediction based imputation. For this, the regression curve is generated and then the regression curve determines the overall value that can be imputed.

Random Forest Approach for missing value imputation

This work is having the key focus on the random based imputation approach[24] [25]is having the fitness score based on the final outcome and overall acceptability score. The proposed work is having mix from random forest approach whereby the division of general dataset is utilized. In division, the total datasets is partitioned to n irregular sections and each fragment is having their own arrangement of missing qualities. These missing qualities are exclusively prepared utilizing the approach is discretionary choice trees utilizing random forest paradigm. The choice tree in each gathering is handled to have the area mean occurrences and hence the outcomes or result of expected mean qualities is assessed from each set. The ascribed an incentive from every choice tree is related with a particular acknowledgment score of positioning and this kind of scoring is utilized to at long last have the last credited esteem which is best fit for attribution.



Specifically for Missing Value Imputation

ISSN: 2229-6913(Print), ISSN: 2320-0332(Online)

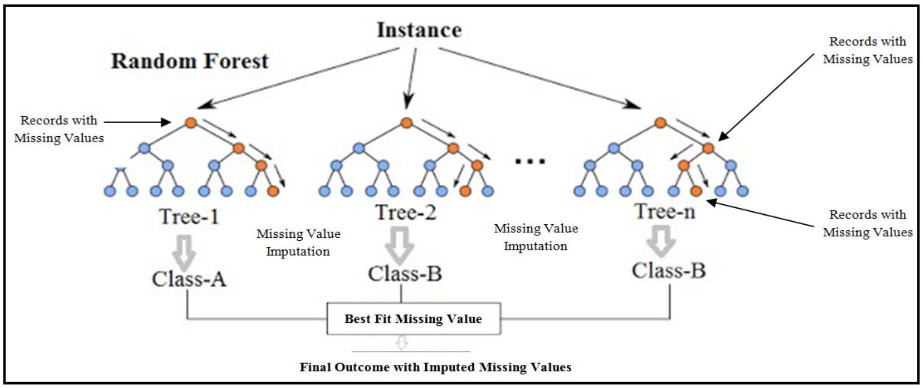


Figure 2 Random Forest Approach with Multiple Decision Tree

Algorithm - Imputation of Missing Value in the Fetched Dataset

Phase-1: Data Extraction and Pre-Processing

1) Extraction of Live Data
2) Identification of Missing Value Attribute to be processed
3) Marking of Missing Values in the extracted record
4) Storage of attribute values to database as well as file system for missing value imputation
5) Generation of possible set of probable missing values to be imputed based on the fitness score
6) Evaluate the Duncan Mean Evaluation in association with fitness approach to associate the best fit strategy
7) Assign and rank the fitness value to each value in the set for imputed set
8) Allocate the value of element to attribute and implement missing value imputation

Phase-2: Imputation using Soft Computing based Approach integrating Random Forest based Best Decision

1) Identification of records with Unidentified or Missing Values and formation of decision tree
2) Fill the incomplete records with the corresponding features of the nearest best fit decision based ranking and threshold evaluation
3) X an n × p matrix, stopping criterion μ with grouping factor
4) Record which variables and which positions have Unidentified or Missing Values in X
5) p_0 number of variables that have Unidentified or Missing Value
6) $X_{FitnessValue}$ based imputation
7) Set ThresholdDifference = ExpectedOptimized while ThresholdDifference > μ do
8) $X_{old.imp}$ $X_{FitnessValue}$
9) Randomly separate the p_0 variables into K = K (alpha) groups of approximately the same size (if = 1, K = p_0)

```
10) for i = 1, ...,K do
11) Let $X_i$ be the columns of X corresponding to group i, $X(_{-i})$ the columns of X excluding group i
12) Set $X_{FitnessValue}$ the values in Xi which were missing back to NA
13) Fit multivariate random forest using variables in groups i as response variables, and the rest as predicting variables. Note: ONLY the non-Unidentified or Missing Values of Xi will be used in calculating the composite split rule
14) $X_{FitnessValue}$ Get the final summary imputed value using the terminal average for continuous variables and using the maximal terminal node class rule for categorical variables
15) end loop
16) Set ThresholdDifference = E ($X_{old.imp}$,$X_{FitnessValue}$)
17) end loop
18) Return the imputed matrix $X_{FitnessValue}$
```

Implementation and Experimental Results

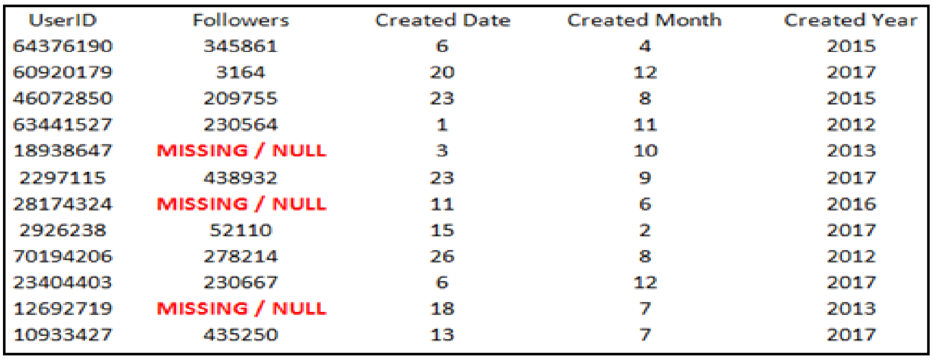
Python Programming is used for implement the random forest imputation approach and data is fetched from different social media using APIs. In this research work on missing value imputation, the situation of removing particular content is taken so the tweets and messages identified with that word can be assessed with missing value attribution. The content or client convictions in light of various classes are taken with the live extraction of information from various web-based social networking. These classes are utilized so the total conclusion about area, instruction, and commonwealth can be broke down with the extraction of information identified with client courses of events. In the greater part of the tweets, the assessment or client convictions of web-based social networking are identified with the academicians and researchers who display their perspectives or conclusions on this online life. The evaluation of followers is done because it is numeric value and identification can be done where there are negligible or null followers. If we do not consider these values, the results can be unpredictive to the final conclusion.

In traditional methodology, such qualities based records are erased however by this approach the ultimate result can't be great. In the separated tweets and messages, it is discovered that the exploration points of view and rules set around various classes are for the most part talked about by the analysts, academicians and government official. With execution of proposed approach, the missing quality imputation is observed to be strong with the filling or attribution of the qualities utilizing random forest approach that is one the unmistakable approach settling on utilization of tremendous choice trees and afterward last positioning of the best result with the scoring of results.

Raw data after preprocessing with missing value

Following screenshot present that the data is collect from the different social media and show the missing or null values after processed it. It shows data in CSV (comma-separated values) format that is readable form. Table 1 depicts that these missing or null value need for imputation because these result create problem for analysis the data.

Table 1 Data with missing value

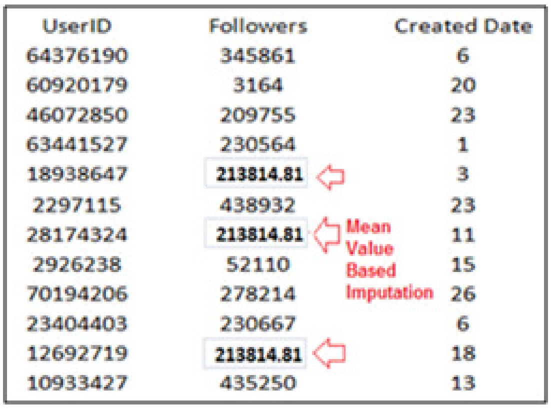


UserID	Followers	Created Date	Created Month	Created Year
64376190	345861	6	4	2015
60920179	3164	20	12	2017
46072850	209755	23	8	2015
63441527	230564	1	11	2012
18938647	MISSING / NULL	3	10	2013
2297115	438932	23	9	2017
28174324	MISSING / NULL	11	6	2016
2926238	52110	15	2	2017
70194206	278214	26	8	2012
23404403	230667	6	12	2017
12692719	MISSING / NULL	18	7	2013
10933427	435250	13	7	2017

Missing Value Imputation using Mean valued based approach

Following is the point of view of execution situation in programming language for missing quality imputation. The table 2 is showing the imputed values using mean imputation.

Table 2 Data with missing value imputation using mean



UserID	Followers	Created Date
64376190	345861	6
60920179	3164	20
46072850	209755	23
63441527	230564	1
18938647	213814.81	3
2297115	438932	23
28174324	213814.81	11
2926238	52110	15
70194206	278214	26
23404403	230667	6
12692719	213814.81	18
10933427	435250	13

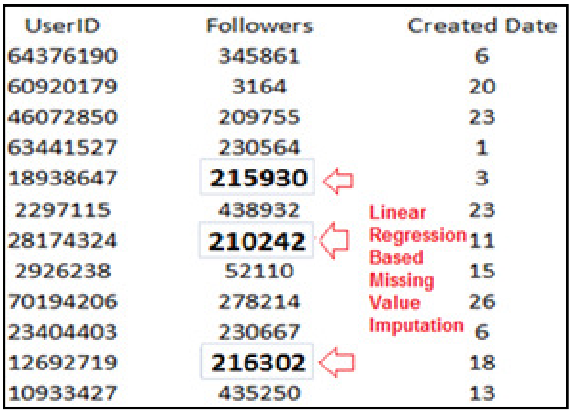
Mean Value Based Imputation

Missing Value Imputation using random based approach

Missing Value Imputation using Linear Regression based approach

The screenshot is introducing the imputed values using Linear Regression based approach. These missing values are different from the mean value imputation

Table 3 Data with missing value imputation using Linear Regression approach

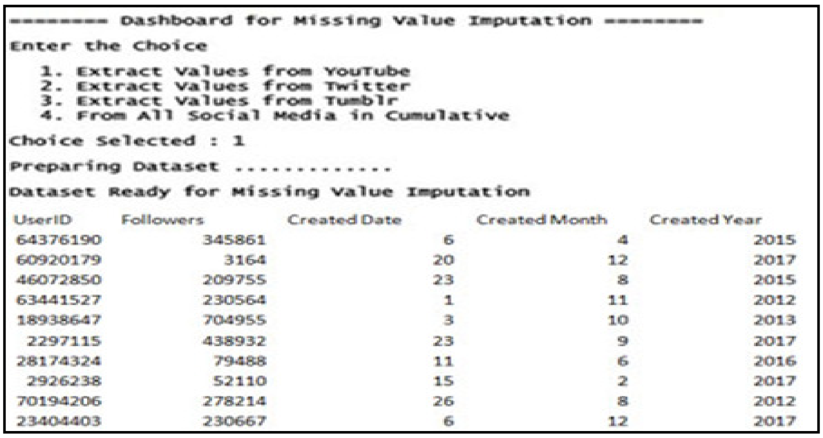


UserID	Followers	Created Date
64376190	345861	6
60920179	3164	20
46072850	209755	23
63441527	230564	1
18938647	215930	3
2297115	438932	23
28174324	210242	11
2926238	52110	15
70194206	278214	26
23404403	230667	6
12692719	216302	18
10933427	435250	13

Linear Regression Based Missing Value Imputation

The screenshot is showing the treated values using random forest based approach of missing value treatment. These imputed values are different from the mean value and the linear regression based imputation.

Table 4 Data with missing value imputation



```
-------- Dashboard for Missing Value Imputation --------
Enter the Choice
    1. Extract Values from YouTube
    2. Extract Values from Twitter
    3. Extract Values from Tumblr
    4. From All Social Media in Cumulative
Choice Selected : 1
Preparing Dataset .............
Dataset Ready for Missing Value Imputation
UserID      Followers     Created Date    Created Month    Created Year
64376190       345861                6                4            2015
60920179         3164               20               12            2017
46072850       209755               23                8            2015
63441527       230564                1               11            2012
18938647       704955                3               10            2013
 2297115       438932               23                9            2017
28174324        79488               11                6            2016
 2926238        52110               15                2            2017
70194206       278214               26                8            2012
23404403       230667                6               12            2017
```

Summary of the Research Variables and Records Identified with Missing Values Extracted from different social media

Following is the analytics of the general executions in exceptional social media in terms of lacking values extracted in unique timelines. The following table 5 presents the missing value obtained from the social media of YouTube, Tumblr and Twitter on the record sets of 200 logs.

Table 5 Social Media related Missing Values

Social	Records Fetched using	Missing

Media	APIs	Values
Twitter	200	11
YouTube	200	15
Tumblr	200	19

The table 5 present that the tumblr is having higher wide variety of missing values as compared to others.

Following table offers the final results of processing time for the missing cost identification on extraordinary social media. As this work is based totally at the analytics of social media based lacking fee imputation, the effects show that the tumblr is having better logs of extra execution time.

Table 6 Evaluation of Processing Time for Missing Value Identification

Social Media	Missing Values	Processing Time in ms for Missing Value Identification
Twitter	11	2.4
YouTube	15	2.8
Tumblr	19	3.1

These results show that Twitter is having minimum processing time in the analytics of missing values.

The following table depicts the time taken by the APIs in the extraction of Tweets with the missing values based on the keyword size and related execution time.

Table 1.6 Time taken by the Social Media APIs in Extraction of Tweets with Missing Values

Social Media Platform	Data Extraction Time (ms)	Keyword Size (No. of Characters)
YouTube	8.6	3
Twitter	10.4	3
Tumblr	9.3	3
YouTube	10.5	5
Twitter	12.4	5
Tumblr	11.4	5
YouTube	10.3	8
Twitter	14.2	8
Tumblr	12.4	8

Following is the styles of data extracted from social media from distinct class and instance with the datasets used on assorted subjects.

Table 1.7 Datasets Evaluation under Classes and Instances

Datasets	Instances	Dataset Size (KB)	Classes
UGC	548	96	1
Gujarat	680	37	2
Disaster	697	48	2
Election	536	60	3
University	673	60	1
City	905	46	2

Table 8 is the depiction of class definitions with the facts evaluation scenarios. Inside the subsequent effects, the accuracy degree is measured.

Table 1.8 Class Definitions and Records

Class	Id	Records Evaluated
Education	1	1219
Location	2	2211
Politics	3	487

Following table 9 is the percentage of accuracy finished after implementation using specific processes of mean, regression and proposed method.

Table 9 Approach based Accuracy

Category	Traditional Mean	Regression Based Outcome	Random Forest
Education	84.48	76.78	95.67
Location	81.39	78.48	94.38
Politics	87.37	79.83	96.34

The accuracy is evaluated with the benchmark of traditional suggest. the evaluation of blunders aspect from the proposed method is evaluated and compared from mean based analysis. the deviation among the values obtained from proposed and classical method is the bottom of the assessment of mistakes factor in absolute and relative terms.

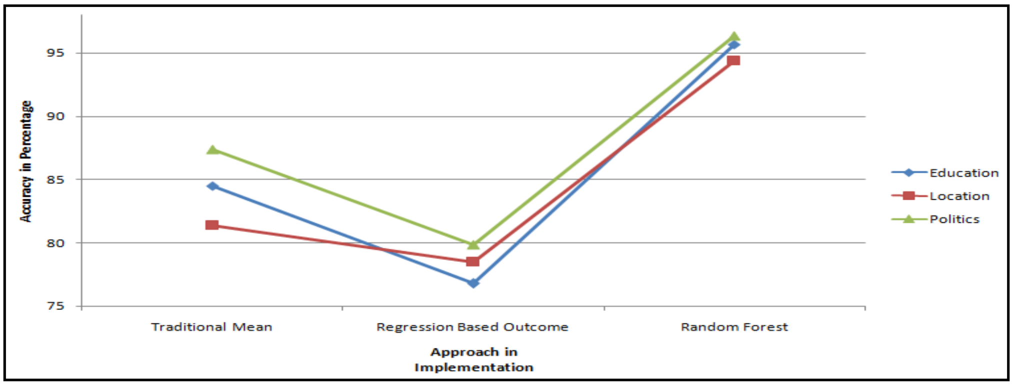


Figure 3 Approach based Accuracy

Comparison of Execution Time in Proposed Approach and Earlier Approaches

Table 5.10 Comparison of Execution Time

	Random Forest based Missing Value Imputation	Classical Mean based Approach	Classical Linear Regression based Missing Value Imputation
Scenario 1	2.7899	2.9888	2.3838
Scenario 2	3.1122	3.2929	2.7888
Scenario 3	3.1383	3.3211	2.2877
Scenario 4	2.8991	2.9878	2.4993
Scenario 5	2.5828	2.6181	2.1738
Scenario 6	2.0711	2.0927	2.0122
Scenario 7	2.0531	2.1929	1.8394
Scenario 8	2.9828	3.2022	2.7889
Scenario 9	2.9292	3.2728	2.7988
Scenario 10	2.8293	3.2333	2.8949

Evaluation and Comparison of Integrity in Proposed Approach and Earlier Approach

Integrity signifies the consistency of the set of rules in terms of jogging in distinctive key phrases. From the effects, it's far obtrusive that the proposed random forest method is integrity and consistency aware on exclusive situations of execution with special keywords in comparison to the preceding processes of imply and linear regression. Following Figure 4 indicates the line graph of consistency and integrity of the results with different key phrases. The effects show that the execution time is integrity aware and constant without any ambiguity.

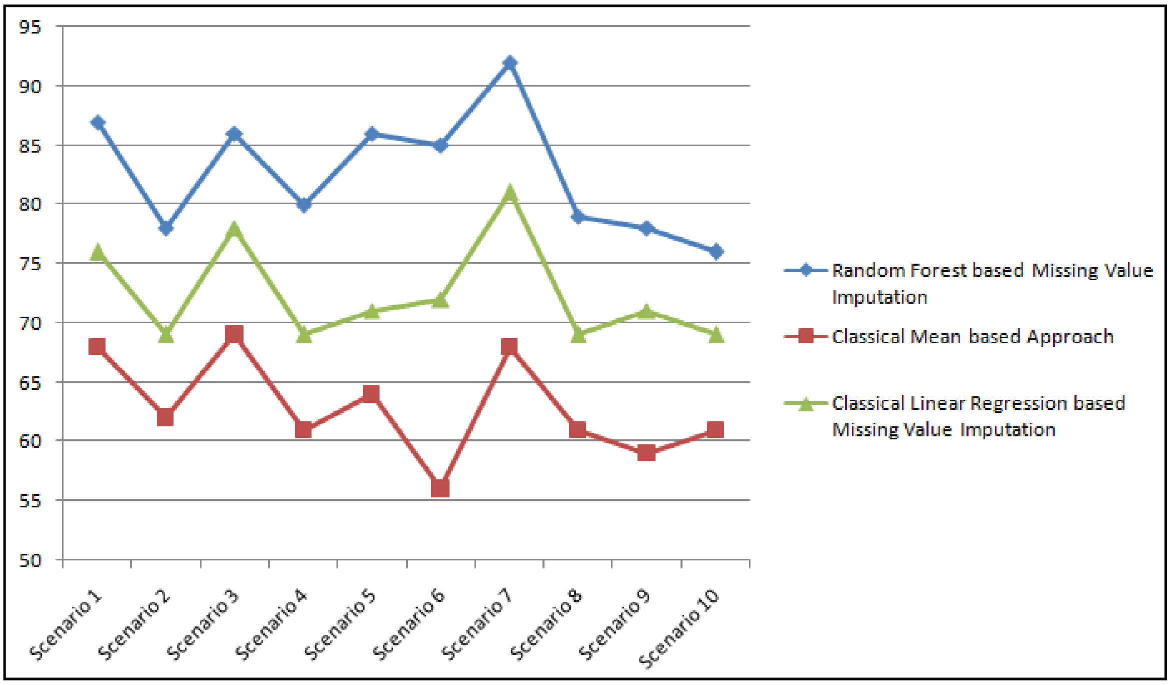


Figure 4 Evaluation of Integrity

Conclusion and Future scope

Using soft computing and machine learning based algorithms the overall effectiveness and outcome can be improved. The proposed approach is not implemented so far in the segment of multiple social media and this is the key focus in this research work. This work extracts the live data from multiple social and extracts the missing values. On extracted missing values, the global fitness score for missing value imputation is done for higher accuracy using random forest approach. There are assorted soft computing and nature inspired approaches which can be further analyzed.

References

[1] Bifet, A., & Frank, E.. Sentiment knowledge discovery in twitter streaming data. In *International Conference on Discovery Science*. Springer Berlin Heidelberg, 2010

[2] Bollen, J., Mao, H., & Pepe, A.. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11, 450-453, 2009

[3] Bollen, J., Mao, H., & Pepe, A.. Determining the Public Mood State by Analysis of Microblogging Posts. In *ALIFE* (pp. 667-668), 2010

[4] Asur, S., & Huberman, B. A.. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 *IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE, 2010

[5] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P.. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1397-1405). ACM, 2011

[6] Saif, H., He, Y., & Alani, H.. Semantic sentiment analysis of twitter. In *International Semantic Web Conference* (pp. 508-524). Springer Berlin Heidelberg, 2012

[7] Leong, C. K., Lee, Y. H., & Mak, W. K.. Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications*, 39(3), 2584-2589, 2012

[8] Wang, H., Cambria, E., Schuller, B., Liu, B., & Havasi, C.. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, 28(2), 12-14, 2013

[9] Dong, H., Shahheidari, S., & Daud, M. N. R. B.. Twitter sentiment mining: A multi domain analysis. In Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 *Seventh International Conference* (pp. 144-149). IEEE, 2013

[10] Cambria, E., Fu, J., Bisio, F., & Poria, S.. AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In *AAAI* (pp. 508-514), 2013

[11] Kotwal, A., Fulari, P., Jadhav, D., & Kad, R.. Improvement in Sentiment Analysis of Twitter Data Using Hadoop. *Imperial Journal of Interdisciplinary Research*, 2(7), 2014

[12] Poria, S. Cambria, E., Fu, J., Bisio, F., &. AffectiveSpace 2: *Enabling Affective Intuition for Concept-Level Sentiment Analysis.* In AAAI (pp. 508-514), 2015

[13] Wehrmann J, Becker W, Cagnini HE, Barros RC. A character-based convolutional neural network for language-agnostic Twitter sentiment analysis. InNeural Networks (IJCNN), 2017 *International Joint Conference on 2017* May 14 (pp. 2384-2391). IEEE,2017

[14] Al-Rubaiee H, Qiu R, Li D. Identifying Mubasher software products through sentiment analysis of Arabic tweets. *Industrial Informatics and Computer Systems (CIICS),* 2016 International Conference on 2016 Mar 13 (pp. 1-6). IEEE, 2016

[15] Heredia B, Khoshgoftaar TM, Prusa J, Crawford M. Cross-domain sentiment analysis: An empirical investigation. *Information Reuse and Integration (IRI),* 2016 IEEE 17th International Conference on 2016 Jul 28 (pp. 160-165). IEEE, 2016

[16] Blaz CC, Becker K. Sentiment analysis in tickets for it support. InMining Software Repositories (MSR), 2016 IEEE/ACM 13th *Working Conference* on 2016 May 14 (pp. 235-246). IEEE, 2016

[17] Fiarni C, Maharani H, Pratama R. Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique. InInformation and Communication Technology (ICoICT), 2016 4th *International Conference* on 2016 May 25 (pp. 1-6). IEEE, 2016

[18] Pamungkas EW, Putri DG. An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia. *Engineering Seminar (InAES),* International Annual 2016 Aug 1 (pp. 28-31). IEEE, 2016

[19] Nithya R, Maheswari D. Correlation of feature score to overall sentiment score for identifying the promising features. *Computer Communication and Informatics (ICCCI),* 2016 International Conference on 2016 Jan 7 (pp. 1-5). IEEE, 2016

[20] Bouazizi M, Ohtsuki TO. A pattern-Based approach for Sarcasm Detection on Twitter. IEEE Access. 2016;4:5477-88, 2016

[21] Gatti L, Guerini M, Turchi M. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*. 2016 Oct 1;7(4):409-21, 2016

[22] Biltawi M, Etaiwi W, Tedmori S, Hudaib A, Awajan A. Sentiment classification techniques for Arabic language: A survey. Information and Communication Systems (ICICS), 2016 *7th International Conference on 2016* Apr 5 (pp. 339-346). IEEE, 2016

[23] Rabab'ah AM, Al-Ayyoub M, Jararweh Y, Al-Kabi MN. Evaluating sentistrength for arabic sentiment analysis. Computer Science and Information Technology (CSIT), 2016 *7th International Conference on 2016* Jul 13 (pp. 1-6). IEEE, 2016

[24] Barve, Abhishek, Manali Rahate, Ayesha Gaikwad, and Priyanka Patil. "Terror Attack Identifier: Classify using KNN, SVM, Random Forest algorithm and alert through messages." *International Research Journal of Engineering and Technology (IRJET)* 2018

[25] Estee, Jan. "Using Machine Learning to Detect Fake Identities: Bots vs Humans". *IEEE Access*. Volume 6 2018

[26] Fengfeng Fan, Zhanhuai Li and Yanyan Wang, "On-Line Imputation for Missing Values", *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics CISP-BMEI 2017*

[27] Bichen Shi, Gevorg Poghosyan, Georgiana Ifrim, and Neil Hurley, "Hashtagger+: Efficient High-Coverage Social Tagging of Streaming News", *IEEE Transactions on Knowledge and Data Engineering, 2018*

[28] Lidong Wang,Randy Jones, "Big Data Analytics for Disparate Data", *American journal of Intelligent System, 2017.*